Establishing the Utility of Non-Text Search for News Video Retrieval with Real World Users

Michael G. Christel
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213 USA
1-412-268-7799

christel@cs.cmu.edu

ABSTRACT

TRECVID participants have enjoyed consistent success using storyboard interfaces for shot-based retrieval, as measured by TRECVID interactive search mean average precision (MAP). However, much is lost by only looking at MAP, and especially by neglecting to bring in representatives of the target user communities to conduct such tasks. This paper reports on the use of within-subjects experiments to reduce subject variability and emphasize the examination of specific video search interface features for their effectiveness in interactive retrieval and user satisfaction. A series of experiments is surveyed to illustrate the gradual realization of getting non-experts to utilize non-textual query features through interface adjustments. Notably, the paper explores the use of the search system by government intelligence analysts, concluding that a variety of search methods are useful for news video retrieval and lead to improved satisfaction. This community, dominated by text search system expertise but still new to video and image search, performed better with and favored a system with image and concept query capabilities over an exclusive text-search system. The user study also found that sports topics mean nothing for this user community and tens of relevant shots collected into the answer set are considered enough to satisfy the information need. Lessons learned from these user interactions are reported, with recommendations on both interface improvements for video retrieval systems and enhancing the ecological validity of video retrieval interface evaluations.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems – *evaluation, video*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.7: Digital Libraries – *user issues*

General Terms

Experimentation, Human Factors

Keywords

Interactive video retrieval, TRECVID analysis, user studies

1. INTRODUCTION

Automated tool support in combination with human manipulation and interpretation offer tremendous leverage in addressing the challenge of video information retrieval. Without automated tools, the human user is swamped with too many possibilities as the quantity and diversity of video accessible on the Web proliferate. Ignoring the human user, though, is a mistake. Fully automated systems involving no human user have consistently and significantly underperformed compared to interactive human-in-the-loop search systems evaluated in the video search tasks of the NIST TREC Video Retrieval evaluation forum (TRECVID) for the past five years [9]. Technology-driven multimedia content indexing approaches have been published with great success rates but merited little or no use in fielded systems, in part because of the often mentioned semantic gap in the multimedia research community, i.e., the lack of coincidence between the information that one can automatically extract from the visual data and the interpretation that the same data has for a user in a given situation [12]. By accounting for and integrating the technological capabilities of multimedia indexing, machine learning, and other applicable techniques with a focus on the human user capabilities and strengths, we can better enable the intelligent human user to efficiently, effectively access relevant video materials from large collections with great satisfaction.

Automated tools, when presented poorly, can lead to frustration or be ignored, but this is not often caught by TRECVID interactive search task reports. TRECVID interactive evaluations historically have emphasized only information retrieval effectiveness measures (e.g., mean average precision – MAP), and not other measures of end user utility. The vast majority of TRECVID interactive runs have been conducted by the researchers themselves posing as users [9], with a few notable exceptions [1, 2, 3, 6, 17]. Even these exceptions, though, use students as subjects rather than real world users. The novelty of this paper is in bringing a suite of HCI methods to bear on the question of video retrieval utility, working with real users to empirically establish the validity of design choices outlined in Section 2. The focus is on an important question facing the video retrieval research community: the value of non-text query. Such query mechanisms are applicable even in the absence of any speech narrative. The paper looks through a series of user studies in Sections 3-5 to go beyond MAP and consider transaction logs and questionnaires as well for measuring use and subjective satisfaction. The paper culminates in a first look at intelligence analysts' participation in a video retrieval experiment addressing the utility of non-text query mechanisms.

MM'07, September 23–28, 2007, Augsburg, Bavaria, Germany.
Copyright 2007 ACM 978-1-59593-701-8/07/0009...$5.00.

2. VIDEO SEARCH: QUERY BY TEXT, BY IMAGE, AND BY CONCEPT

Today's commercial video search engines often rely on filename and accompanying text sources [13]. Users issue text queries to retrieve nonlinguistic visual imagery. The image retrieval community has focused on content-based indexed by pixel-level image attributes like color, texture, and shape [12, 13], where users supply a visual example as the search key, but the underlying low-level attributes makes it difficult for the user to formulate queries. In an attempt at bridging this semantic gap, the multimedia research community has invested in developing a Large-Scale Concept Ontology for Multimedia (LSCOM), whereby semantic concepts like "road" or "people" can be used for video retrieval [8]. These three access strategies, query-by-text, query-by-image example, and query-by-concept, can be used to produce storyboard layouts of imagery matching the issued query. Through the past five years, interactive retrieval systems evaluated in TRECVID have almost universally supported query-by-text, with that functionality responsible for most of the information retrieval success through TRECVID 2004 [5]. Query-by-image example is the next most frequently supported strategy across TRECVID participants [4, 5, 9], with query-by-concept not having success in early 2003-2004 trials [6, 17] and not being implemented and tested as widely as the other query strategies.

All three strategies (query by text, image, concept) have been used to produce storyboard layouts of imagery by the Carnegie Mellon Informedia video search engine [1, 2, 3] and the MediaMill video search engine [13, 14] for a number of years, with these systems scoring best for all of the TRECVID interactive video search evaluations since the task inception in 2002 [9]. Hence, there is evidence that the three strategies together are effective for the TRECVID search tasks, but there is a qualification. Those top-scoring runs have consistently been produced by **"expert"** runs, with a focus of the user studies research reported here being an understanding of **"novice"** users' activity. The expert runs establish idealistic upper bounds on performance, at the expense of assuming certain knowledge and motivation by the expert users. Throughout this paper we will use the term "expert" to refer to a user with three sources of knowledge not possessed by "novices": (1) the expert has been working with the research group for at least a year, having a better sense of the accuracy of various automated video processing techniques; (2) the expert has used the tested video retrieval system prior to timed runs with the TRECVID data, perhaps even contributing to its development, and therefore knows the system operation better than study participants who first see it during the test run; and (3) the expert knows about TRECVID evaluation, e.g., the emphasis on shot-based retrieval and use of mean average precision as a key metric. The focus of this paper is understanding the utility of query-by-image and query-by-concept for novices who have no implicit motivation to score well according to standard TRECVID metrics and who are using the given video access system for the first time.

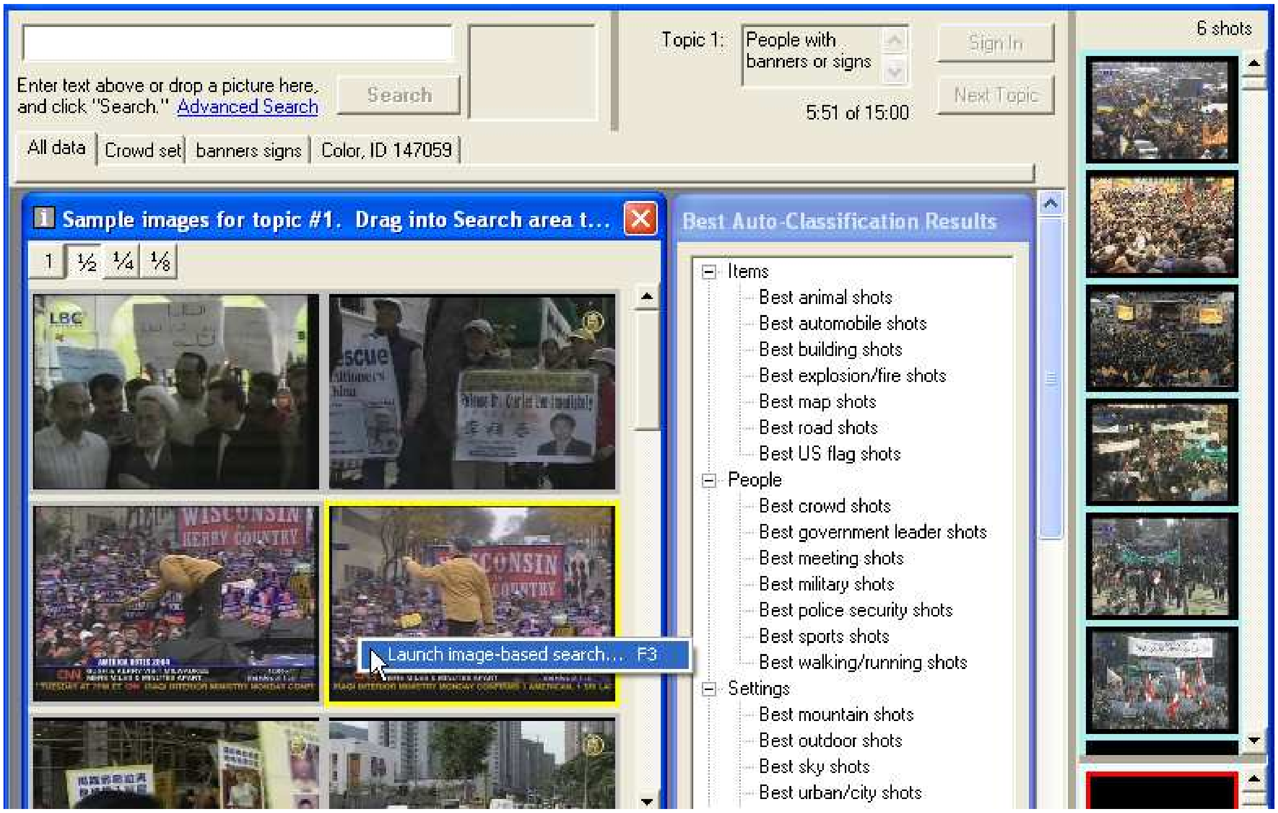
In video processing, a broadcast is commonly decomposed into numerous shots, with each shot represented by a keyframe: a single bitmap image extracted from that shot. The numerous keyframes can then be subjected to image retrieval strategies. This simplified approach to video retrieval is the focus in the studies reported here, with the benefit that many of the lessons learned for such shot-based video retrieval will be applicable as well for still image retrieval.

TRECVID at NIST is an evaluation forum with an interactive search task measuring the effectiveness of shot-based retrieval. The TRECVID search task is defined as follows: given a multimedia statement of information need (topic) and the common shot reference, return a ranked list of up to 1000 shots from the reference which best satisfy the need. Success is measured based on quantities of relevant shots retrieved in the set of 1000, in particular the metrics of recall and precision. The two are combined into a single measure of performance, average precision, which measures precision after each relevant shot is retrieved for a given topic. Average precision is then itself averaged over all of the topics to produce a mean average precision (MAP) metric for evaluating a system's performance [9]. There are 23 graded topics for TRECVID 2004, working against 64 hours (128 broadcasts) of ABC News and CNN Headline News video from 1998, consisting of 33,367 reference shots. There are 24 graded topics for TRECVID 2005, working against 85 hours (140 international broadcasts) of English language, Arabic, and Chinese news from 2004, consisting of 45,765 reference shots.

3. BASELINE SYSTEM (2004): NOVICES IGNORING IMAGE/CONCEPT QUERY

In 2004 a user study was conducted with 24 university students and staff who had no familiarity with TRECVID or the interface being investigated, and no connection to the research team – 'novices" according to our definition in Section 2. We created two systems with nearly identical user interfaces and search capabilities, but with one system, Visual-Only, being a system ignorant of the speech narrative. The other system, Full, had access to the speech narrative metadata as well as all visual metadata. Both systems provided query-by-text, query-by-image (color-based), and query-by-concept (8 concepts) functionality, but for Visual-Only, the query-by-text only worked against recognized on-screen display text, not speech transcripts. Participants answered 2 TRECVID 2004 topics in each system, and filled out accompanying questionnaires.

Participants using the Full system, with access to the closed-caption and ASR transcripts, score significantly higher on the performance metric of average precision. The questionnaire data shows the participants had an overwhelming preference for Full over Visual-Only, despite their nearly identical appearance [1]. Transaction log analysis shows that users with a "visual-only" system still relied heavily on text search against visually presented text. The interface encouraged the use of text search over image search and concept browsing because the latter two required extra steps and interpretation. We also had expert use of the same Full and Visual-Only systems to compare against novice use. The expert runs outperform the novice runs with the respective systems. The data in the transaction logs of experts and novices shows the experts made use of query-by-text less and visual strategies (query-by-image, query-by-concept) more, but the novices relied heavily on text search, despite poor quality and poor coverage of news with automatically detected on-screen text within the Visual-Only treatment [1].


Enter text above or drop a picture here, and click "Search."

Topic 1: People with banners or signs

5:51 of 15:00

Figure 1. Interactive search interface for user studies, with query-by-text (top left), query-by-image (middle left), query-by-concept (middle), topic description (top middle), and collected answer set display (right) all equally accessible. The design guideline "what happened and why" is addressed by communicating what type of query is being issued and augmenting results with text and image captions and other feedback mechanisms, e.g., a look within the "Crowd set" results tab would show added detail describing the storyboard of shot imagery being the result from a query-by-concept, specifically "Crowd", along with a "Crowd" description.

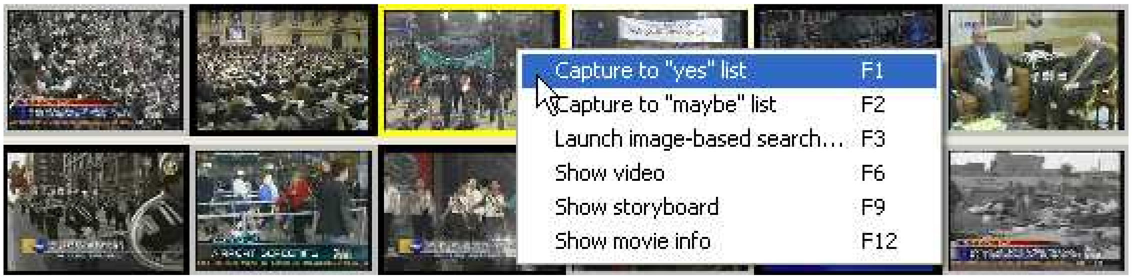


Figure 2. Context-sensitive menu of shot-based actions available for all thumbnail representations in the interface.

Regardless of treatment, Full or Visual-Only, text search strategies dominated, as found by other TRECVID search investigators (e.g., [6, 17]), but the excellent performance of experts when using all three query strategies led us to the goal of enabling increased use of these effective query mechanisms by novices as well.

Making use of published design guidelines for clarifying search interfaces [10], the system's interface was redesigned to improve the user's understanding of "what happened and why" and to promote query-by-text, query-by-image, and query-by-concept equally for novice users. The resulting interface, shown in Figure 1, was then tested in follow-up studies in 2005 (Section 4) and 2006 (Section 5), conducted with TRECVID 2005 topics and different user groups. Later tables and figures summarize the interaction data collectively.

Again based on consistency and clarity recommendations [10], the most common actions were provided in a context-sensitive menu available with all thumbnails (consistency), with the actions labeled on the keyboard (clarity) for easy access through either keyboard or mouse input. These actions are shown in Figure 2. Thumbnails showing a key frame per shot arranged in temporal order in storyboards are typical of most TRECVID search systems (e.g., see [1, 4, 6, 9]), with some extensions. For example, FX Palo Alto uses variable size thumbnails in a collage presentation

[4] rather than fixed size, and MediaMill provides a "cross browser" for fast browsing of thumbnails in ranked results or timeline order in addition to a grid-based storyboard [14]. Here, we make use only of the simple grid layout as in Figure 2, to focus on the effects of the 3 query strategies in the retrieval interface.

4. REVISED SYSTEM (2005): STUDENT USE OF QUERY BY IMAGE/CONCEPT

In a study conducted in September 2005, 24 university students addressed 4 TRECVID 2005 topics each, in a within subjects experiment such that all 4 topics were presented in the standard Full interface treatment as illustrated in Figures 1 and 2, but with an aggressive user interaction history mining strategy used for 2 of the topics. That aspect of the study is detailed elsewhere and found the mining strategy led to no performance differences [2]. Here we focus on the use of the query-by-text, query-by-image, and query-by-concept (39 LSCOM-lite concepts [8]) system. As for such interactions, university students now made use of query-by-image and query-by-concept strategies, a difference from the 2004 results. Such use was productive and in greater agreement with the expert's interactions: Table 2 and Figure 6 in later sections summarize the interaction logs. The students performed well: the MAP for the 4 runs through the 24 topics ranged from 0.253 to 0.286, the highest MAP for TRECVID interactive search conducted by users outside of the system development teams [2].

While this user study provided empirical data on the use and utility of the different query mechanisms for video shot-based retrieval, it left unanswered the question of whether such search behavior would be typical for the class of users mining video archives for information. Are university students good surrogate representatives of intelligence analysts, or do analysts expect and work better with a different set of tools? Also, if query-by-text, the prevalent strategy for video access now on the Web, is the only strategy available, does usability suffer? These questions are explored in detail for the remainder of the paper.

5. WITHIN SUBJECTS TRECVID STUDY (2006): INTELLIGENCE ANALYSTS

A user study, conducted in September 2006 against TRECVID 2005 topics and detailed here for the first time, had two goals: (1) confirm that intelligence analysts, like the students in the September 2005 study, made use of all provided strategies with a resulting good performance on tasks (the desire to compare back against the September 2005 study led to the use of the same 24 TRECVID 2005 topics); (2) through a within-subjects experiment including transaction logs and questionnaires, quantify and qualify the differences between simplified multimedia retrieval systems where only keyword text search is provided, versus the full-featured system offering query-by-text, query-by-image, and query-by-concept. Whereas earlier work confirmed that narrative text was useful for TRECVID shot retrieval (see Section 3), this experiment investigates whether narrative text is sufficient or if the visual concepts and search mechanisms offer benefit, too. The query-by-concept functionality was enriched over that used in the September 2005 study by including the topic-dependent concept set produced through fully automated search against the given topic using statistical machine learning techniques [16]. For example, when given a topic "Tony Blair", a fully automatic process runs to compute the best "Tony Blair" shots, with this shot set then available to the user alongside the other query-by-concept sets for best road, building, etc., shots, i.e., alongside the best shots for the 39 LSCOM-lite concepts [8].

5.1 Participants

Six intelligence analysts were recruited to participate in the experiment as representatives of a user pool for news corpora: people mining open broadcast sources for information as their profession. These analysts (5 male, 1 female), compared to the university students participating in the prior reported studies, were older (2 older than 40, 3 in their 30s, 1 in 20s), more familiar with TV news, just as experienced with web search systems and frequent web searchers, but less experienced digital video searchers. Their expertise was in mining text sources and text-based information retrieval rather than video search. They had no prior experience with the interface under study or data under study and no connection with the research group conducting the study or the NIST TRECVID community. Of course working with even more analysts would have been desirable to better represent the user pool. Global political situations, demands on analysts' time, and logistics limited our access to six individuals over a two-day period.

5.2 Procedure

Participants worked individually with an Intel® Pentium® 4 class machine, medium resolution 1280 x 1024 pixel 18-inch LCD monitor, and headphones. Participants' keystrokes and mouse actions were logged within the retrieval system during the session. They first signed a consent form and filled out a questionnaire about their experience and background. They were then given a paper-based tutorial explaining the various features of the system, and 15 minutes of hands-on use to explore the system, and try the examples given in the tutorial, before starting on the first topic.

We created two systems with nearly identical user interfaces as shown in part in Figures 1 and 2 but with one system, Text-Only, being a "text-only" system making use of only the speech narrative for query-by-text. In the Full system only, query-by-image color similarity search and query-by-concept search (using the 39 LSCOM-lite concepts and topic-dependent concept set) were available. The topics and systems were counter-balanced so that in a first session with 4 topics, the first 2 topics were given as Text-Only or Full and the second 2 topics in the other system, with the analysts each working through a second session of 4 topics in which the system order was reversed. For each topic, the user spent exactly 15 minutes with the system answering the topic, followed by a questionnaire. The questionnaire content was the same as used by all of the TRECVID 2004 interactive search participants, designed based on prior work conducted as part of the TREC Interactive track for several years [9]. The analysts were not told of the details of "Text-Only" and "Full" system variants. Instead, questionnaires referred to the first system used in a session as "System 1" and the second as "System 2." Participants took two additional post-session questionnaires after the fourth and eighth topics for reflections on their System 1 vs. System 2 experiences.

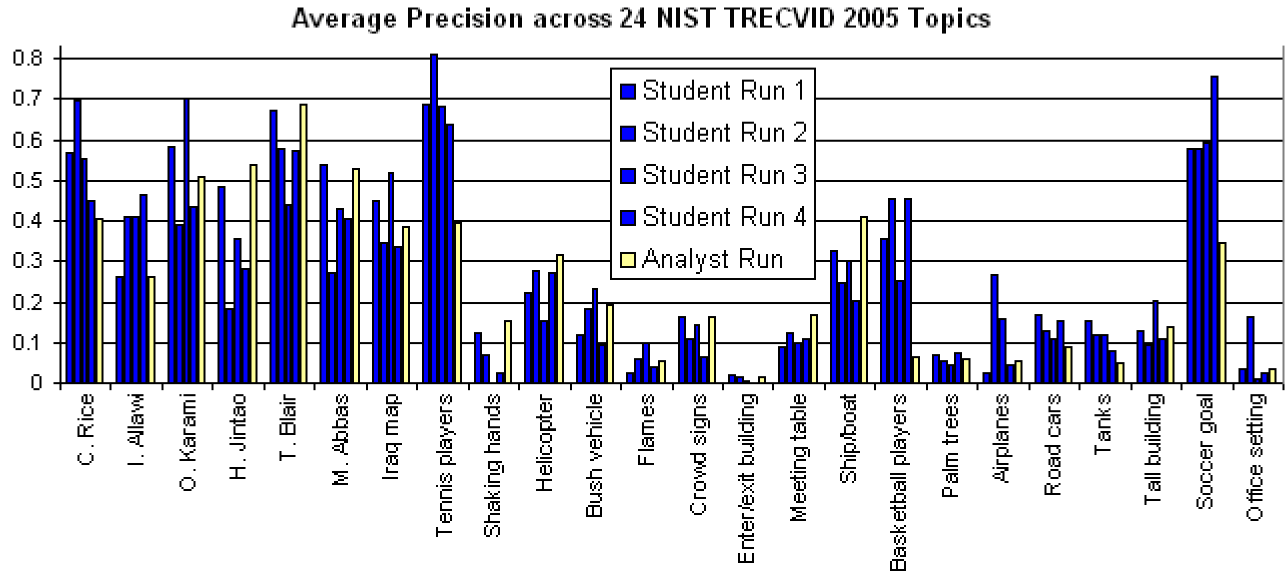


Figure 3. Average Precision across 24 NIST TRECVID 2005 topics, "Full" treatment runs from user studies reported in Section 4 and Section 5. For three sports topics (tennis, basketball, soccer), students do well while analysts underperform.

During a 15-minute topic run, the participant was able to type a query to search against text metadata for video consisting of closed-captions, automatic speech-recognized (ASR) text, machine translations of ASR text, and automatic capture of on-screen text. For the Full treatment only, the user could also select any thumbnail and use it to launch a color-based image query, or select a concept from a text explorer tree-based view of concepts, e.g., select "buildings" to launch a query-by-concept and load the top-ranked 1000 automatically judged building shots. Whatever query was issued, the results would then be posted in a tab within the interface with appropriate counts and labels letting the user know how many items were retrieved and the nature of the query. The results were by default displayed in a scrollable storyboard grid filling the screen with one-quarter resolution thumbnail images, one image per shot. MPEG-1 videos of 352 x 240 pixel resolution were hence by default represented with a series of 88 x 80 pixel thumbnail images. Figure 2 shows a portion of a storyboard – just two rows of six thumbnails each. The default storyboard size was 10 rows of 12 thumbnails each, allowing 120 shots to be viewed at once with scroll support to show additional shots. The user had control over resizing storyboards to change the row and column count and thumbnail size, and could blow-up any thumbnail to its full resolution of 352 x 240 by pressing the Shift key when hovering over the thumbnail.

The participant was in complete control over allocating time between issuing more queries (to generate more tabs), versus exploring a resulting storyboard (contents of a tab) carefully or completely. The time counter (shown as "5:51 of 15:00" in Figure 1) turned red as a warning when less than a minute remained, with the system locking out all user access at 15 minutes.

5.3 Results

The analysts scored well on the TRECVID 2005 topics, especially since the six analysts reported no prior experience at all with video search systems. Their mean average precision (MAP) of 0.251 when using the Full system correlates well with the 4 student runs' MAP in the study of Section 4 of 0.253 through 0.286. Looking at the average precision across the 24 topics shown in Figure 3, the analysts underperformed compared to the students on three "easy" tasks where the students performed well: topics 8 ("tennis players"), 17 ("basketball players") and 23 ("soccer goal"), the three sports topics. In questionnaire data and later discussions, the analysts indicated disdain and perceived irrelevance for these sports-centered topics as they did not correlate well with their work, so it is not surprising to find that the analysts did not take answering these topics as seriously as the others. If the three sports-related topics are ignored, the MAP for the four student runs of Section 4 are 0.249, 0.228, 0.242, and 0.201, with the analyst run having a MAP of 0.248.

The MAP across all 24 topics for Text-Only was 0.204 while the MAP for Full was significantly better at 0.251 (t=1.87 with 23 df, $p < 0.04$). Removing the 3 sports topics for which the analysts did not put forth a serious effort shows an even more significant difference. The MAP for the 21 non-sports topics for Text-Only was 0.178 while the MAP for Full was 0.248 (t=2.94 with 20 df, $p < 0.005$). The average precision across the 24 topics for the two systems used by the analysts is shown in Figure 4.

The qualitative questionnaires showed that Full not only outperformed Text-Only but was also strongly preferred over it. Each of the 6 analysts participated in 2 sessions of 4 topics each. Regardless of which system was seen first, for 11 of the 12 sessions, Full was noted on the questionnaire as the preferred system, with the analyst for the remaining session indicating no preference for "System 1 vs. System 2" because the choice was topic-dependent (and indeed, that analyst had difficult topics for Full and relatively easy ones for Text-Only in that session). Somewhat surprisingly, the analysts also chose Full over Text-Only as being easier to learn to use and easier to use, despite Full having additional features.

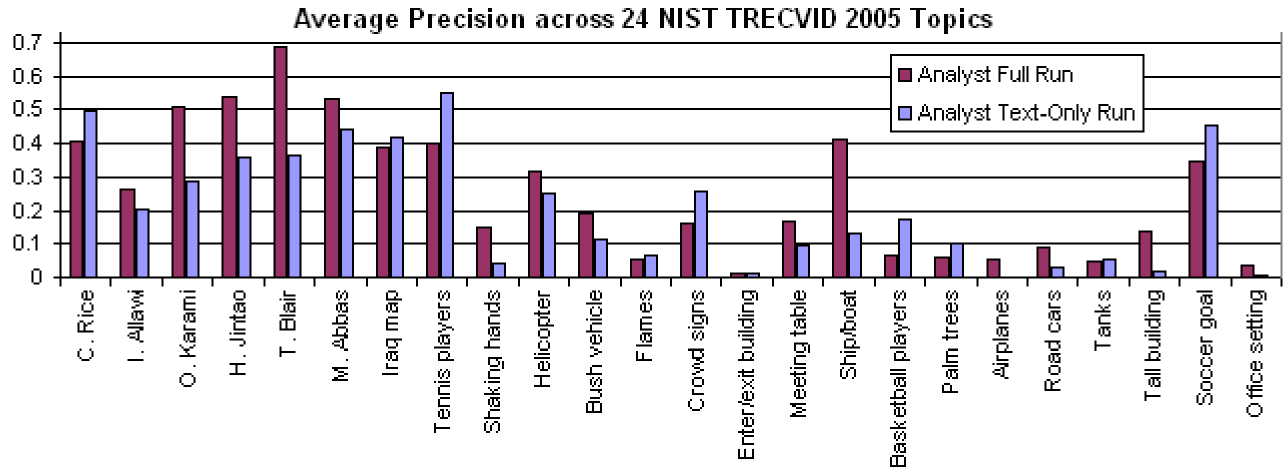


Figure 4. Average precision for analysts addressing TRECVID 2005 topics, Full and Text-Only systems.

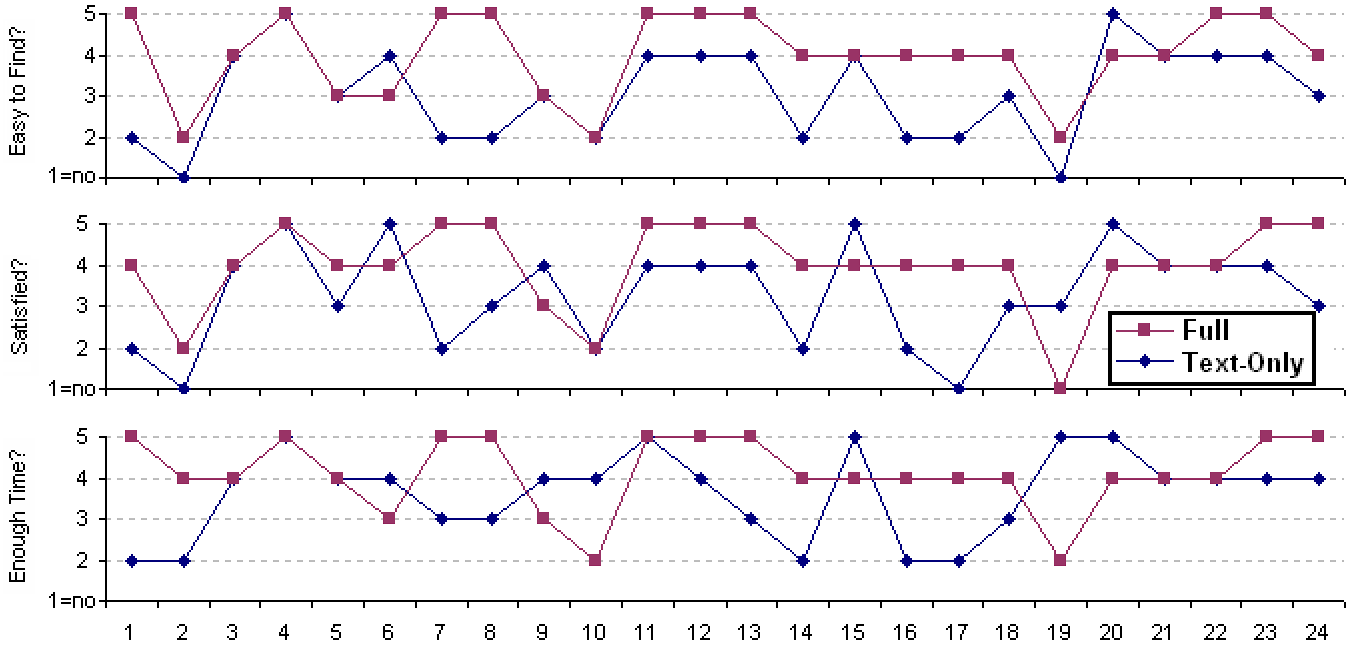


Figure 5. Post-topic questionnaire responses, same TRECVID 2005 topic order as Figs. 3 and 4, with significant differences showing analysts felt that topics were easier to find and produced results more satisfying when using Full system vs. Text-Only.

The post-session questionnaire responses confirm that the streamlined interface redesign, made based on the 2004 user study summarized in Section 3 to better support query-by-image and query-by-concept, indeed works for real world users: the added interface features do not introduce added complexity, with 4 of 5 sessions marked with Full as easier to learn than Text-Only (7 indicating "no difference"), and 7 of 7 sessions marked with Full as easier to use than Text-Only (5 indicating no difference).

Additional questionnaire responses further support the conclusion that the full-featured system was seen as easier to use and resulting in more satisfying performance over a traditional text-only retrieval system. The post-topic questionnaire, answered immediately after each of the 24 topics were completed by each user, contained 5 statements utilizing a 5-point scale with 1=*not at all* and 5=*very much*. For the first two statements – "I was familiar with this topic before I did the search," and "The example images/videos given with this topic description were useful for searching" – there was no significant difference between the Text-Only and Full systems, as anticipated. However, participants responded to the statement "I found that **it was easy to find shots** that are relevant for this topic" very differently: Full systems received a mean rating of 4.00, whereas the Text-Only system received a significantly lower mean rating of 3.08 (t=3.94 with 23 df, $p < 0.0005$). With respect to the statement "For this particular topic **I was satisfied** with the results of my search," again, there was a significant difference (t = 2.6, $p < 0.01$): following the use of Full, the mean rating for "satisfied" was 4.00 compared to 3.29 following the use of Text-Only. The fifth question, "For this topic **I had enough time** to find enough answer shots," was not answered differently at the $p < 0.05$ level of significance. Following the use of Full, the mean rating for "enough time" was 4.125, compared to 3.625 following the use of Text-Only. The responses for "enough time" were dominated by responses of 4 on the 5-point scale, with later interviews confirming that the analysts were not stressed by the 15-minute time limit and generally felt comfortable with their own perceived efficiency. To see what the analysts did with their 15 minutes, we turn to the interaction logs. The topic-by-topic breakdown of these question answers is shown in Figure 5.

The storyboard interfaces allowed for impressive numbers of shots to be reviewed interactively by users within TRECVID's 15-minute time limit per topic. Shots could be judged as correct and put into the "yes" pool posted to the upper right panel answer set shown in Figure 1. They could be judged as "maybe", i.e., possibly correct, and posted to a lower right panel maybe set. Finally, they could be passed over and not judged at all, either dismissed by accident or because they are not relevant to the current topic. Table 1 shows a breakdown of the shot counts on average for the topics from 4 pools: an expert with the Full interface, and then 3 novice pools: students with Full interface (Section 4), analysts with Text-Only, and analysts with Full.

Analyzing the answer sets using NIST pooled truth results, we can see the difference between expert and novice behavior: the expert was motivated to place only very precise, very likely relevant shots into the "yes" set, with more use of the "maybe" set and a bit less exploration of the shot set as evidenced by a smaller overlooked shot count. The students were motivated nearly to the same levels of "yes" shot counts as the expert, but with less precision, 74% later judged correct vs. 93% of the expert's "yes" shots. Details on student vs. expert performance are given elsewhere [2], but the point of interest here is folding in the analyst performances with the two treatments.

Follow-up interviews and questionnaire data show that the analysts were satisfied with their answer sets, felt they had enough time to find the answers, and believed the system worked well in finding relevant shots. These responses were given with the analysts filling out on average only 30.2 shots per topic in their answer set with the Full system, and much less with the Text-Only system. For analysts, finding 30 relevant shots for a topic is enough volume to satisfy the topic, whereas the TRECVID metric looking at finding 1000 relevant shots for a topic is not a good fit to their open broadcast mining and reporting needs. The analysts did not feel compelled to find hundreds of relevant shots, even with the instructions to "find as many as possible in the 15 minutes." They felt that the tens of shots already collected were sufficient to fulfill the task and were satisfied with their relatively small answer sets.

Table 1. Average shot counts and percentage of correct shots per topic for "yes" set, "maybe" (M) set, and overlooked (Skip) set, addressing TRECVID 2005 topics using interface shown in part in Figures 1 and 2.

First 3 rows have query-by-{text, image, concept} functionality	Shot Counts			% Correct		
	Yes	M	Skip	Yes	M	Skip
Expert	58.6	12.0	433.2	92.7	72.1	7.0
Students (Section 4)	52.1	4.2	580.4	74.2	31.8	4.6
Analysts (Full)	30.2	6.2	738.7	77.6	57.7	4.8
Analysts, Text-Only	15.7	5.7	605.1	78.8	44.9	4.8

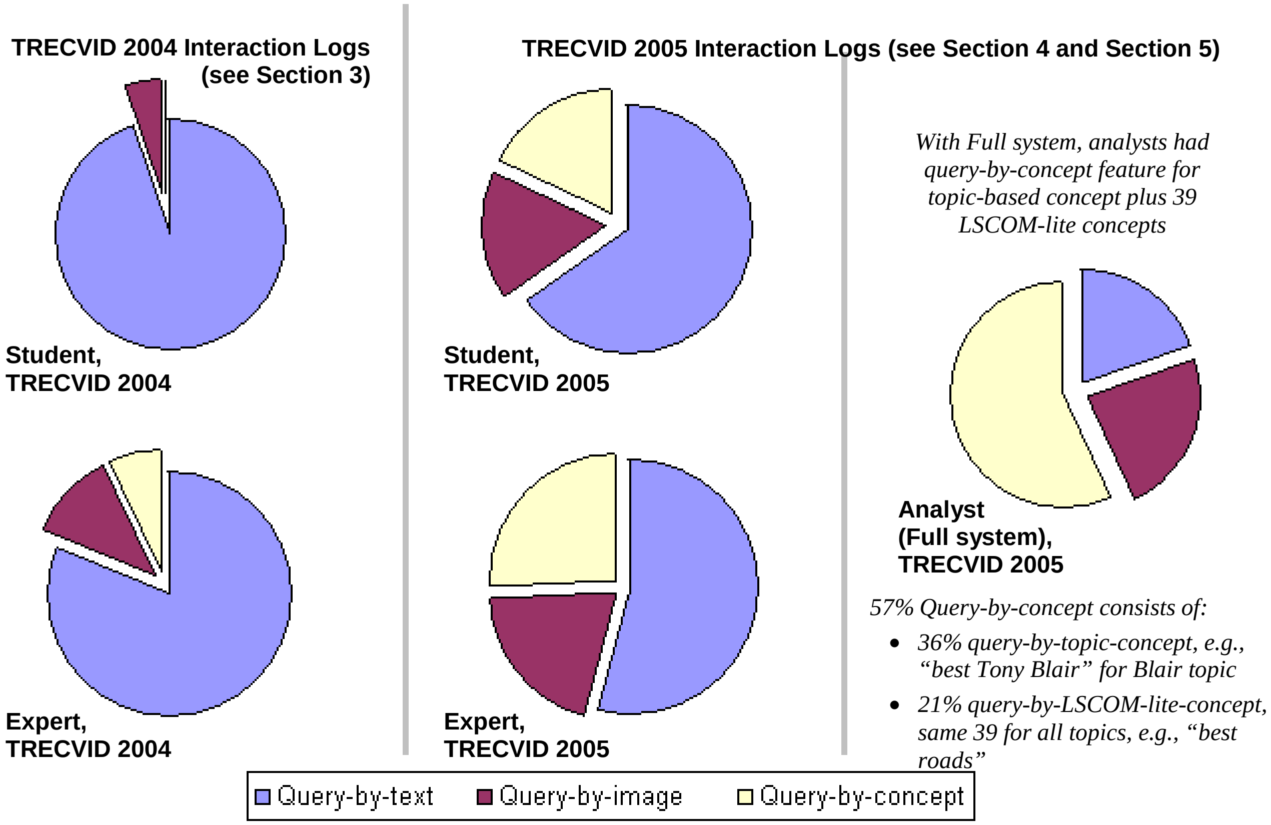
6. DISCUSSION

6.1 Transaction Logs

Table 2 shows the interaction log statistics for students in the 2004 experiment (Section 3) and 2005 experiment (Section 4), and the newly reported experiments with analysts (Section 5). The interface design can clearly affect novice user interaction. A poor interface can deflate the use of potentially valuable interface mechanisms, while informing the user as to what search variants are possible and promoting visibility of system status and recognition over recall – the advice of [10] leading to the design illustrated in part in Figures 1 and 2, can produce a richer, more profitable set of user interactions. The video retrieval interface used in the TRECVID 2005 experiments succeeded in promoting the use of concept search and image search nearly to the levels of success achieved by an expert user, closing the gulf between novice and expert interactions witnessed with a TRECVID 2004 experiment.

Table 2. Summary statistics from novice interaction logs: 2006 study with analysts, TRECVID 2005 (see Section 5) and 2005 study with students, TRECVID 2005 (see Section 4) compared to baseline system with students, TRECVID 2004 (see Section 3).

	Redesigned System (see Figs. 1, 2), TRECVID 2005			Students, *Full* TRECVID 2004 System
	Analysts, *Full* System	Analysts, *Text-Only* System	Students, *Full* System	
Number of users	6	6	24	*24*
Number of topics	24	24	48	*48*
Avg. (average) query-by-concept per topic	1.96	n/a	1.13	*0.13*
Avg. query-by-image per topic	1.75	n/a	4.19	*1.23*
Avg. text queries per topic	3.5	7.29	7.21	*9.04*
Word count per text query	2.8	3.49	2.19	*1.51*
Avg. number of video segments returned by each text query	230.8	165.7	196.8	*105.3*
Query/browse actions per topic	7.21	7.29	12.53	*10.4*


TRECVID 2004 Interaction Logs (see Section 3)

TRECVID 2005 Interaction Logs (see Section 4 and Section 5)

With Full system, analysts had query-by-concept feature for topic-based concept plus 39 LSCOM-lite concepts

57% Query-by-concept consists of:

- *36% query-by-topic-concept, e.g., "best Tony Blair" for Blair topic*
- *21% query-by-LSCOM-lite-concept, same 39 for all topics, e.g., "best roads"*

Figure 6. Percentage of shots submitted via 3 video search strategies, taken from transaction logs.

Figure 6 reflects the changes in system use for gathering shots addressing a topic: for the 2004 system, the novice student users contributed 95% of their shots through query-by-text, with almost all of the rest coming from query-by-image and very few from query-by-concept. For the same TRECVID 2004 topics, the expert contributed 81% of the shots from query-by-text, 12% from query-by-image, and 7% from query-by-concept. For 2005 (Section 4), the novice student users contributed 65% of their shots from query-by-text, 17% from query-by-image, and 18% from query-by-concept. The expert user contributed 54%, 20%, and 26% respectively from these same query sources. In September 2006, the analysts with the Full system contributed 20%, 23%, and 57% from these sources.

From Figure 6, it can be seen that the analysts made significantly more use of query-by-concept than any prior user group and system, despite being proficient in query-by-text interactions. Their greater use can in part be attributed to improved accuracy through machine learning approaches in deriving shot sets for both LSCOM-lite visual concepts like "roads" and in automated topic-based shot sets for stated topics like "Tony Blair" [16]. Also, the greater use of query-by-concept for all groups from TRECVID 2004 to 2005 is attributable to the increased number of concepts in the query-by-concept set from 8 to 39, an argument for more concepts in agreement with the MediaMill finding that 101 concepts led to improved video search performance [13]. The full LSCOM design addresses both human and technology requirements for success, looking for concepts that are observable (related to target video set), feasible (capable of being automated), relevant to genuine use cases and queries, and that cover the overall semantic space of interest to end users [8].

6.2 Design Implications

Think-aloud protocols, transaction logs, interviews, and questionnaires were used with the analysts to solicit additional feedback. The analysts expressed a belief in the potential of query-by-concept functionality (and, indeed, made use of it), but also indicated a sense of frustration with having too many concept options at hand for the task and requested help in making concept selection and concept filtering easier, even with only the 39 concept LSCOM-lite set in place. When the thousand concept LSCOM set is put into the interface, the novice user will need even more help in identifying the potentially useful concepts for addressing an information need.

Returning once more to Figure 6, the analysts trusted and made use of the query-by-topic-concept 36% of the time, and in fact, they used the automatically generated query-by-topic-concept set too much for some topics, failing to leave it for other query strategies. For example, they fell into linear browsing of a shot set like "best Iraq maps" for the "find Iraq maps with Baghdad shown" topic, without ever issuing additional text queries, image queries, or other concept queries. They may have covered more shots with such a straightforward linear browsing strategy, but their precision decreased (as evidenced by Table 1), their submitted shot count decreased (Table 1), and they were too passive at times to take charge and initiate their own queries that could have been more profitable. Future video search systems will likely have these capabilities to promote query-by-topic-concepts when it has potential, and to encourage other query strategies as well to round out searches for relevant material:

- Determine appropriate levels of filtering and use of a large set of concepts, like LSCOM.
- When given a topic, automatically recommend particular concepts for user-driven filtering, i.e., reduce and simplify the options presented to the user based on topic context. For example, for a topic dealing with vehicles, promote the use of road and automobile concepts and suppress animals.
- Present automated use of concept filter combinations for user feedback, with the shot sets produced likely to attract user interactions (with analysts using the Full system, such topic-based shot sets accounted for 36% of the interactions).
- Account for the human user in the search loop by adapting the determination, recommendation, and presentation of query and concept options to the particular user's expertise with the task, the corpus, the system and its concepts. The analysts repeatedly emphasized in their interviews the wish for video retrieval functionality adapted to their personal expertise.

When attributes of a target user community are known, such as the text search expertise of intelligence analysts, the interface should be tuned to work as a better tool leveraging from that expertise and user expectations. For example, the six analysts all assumed the existence of a state-of-the-art text search interface, so when "simple" things like Baghdad spelling correction for "Bagdad" or "Bahgdad" was not provided, they were confused and annoyed. While focusing on query-by-concept in this paper, the basics of the storyboard presentation and simple, efficient layout of the rest of the interface shown in Figures 1 and 2 should not be compromised. The recommendations from [3] based on [10] should be carefully considered, to "capture user interaction history", "provide consistent features", "enable informed choices", and "facilitate efficient investigations."

6.3 TRECVID Implications and Limitations

The NIST TRECVID organizers are clearly cognizant of issues of ecological validity: the extent to which the context of a user study matches the context of actual use of a system, such that it is reasonable to suppose that the results of the study are representative of actual usage and that the differences in context are unlikely to impact the conclusions drawn. TRECVID organizers design interactive retrieval topics to reflect many of the various sorts of queries real users pose, based on query logs against video corpora like the BBC Archives and other empirical data [9]. The topics include requests for specific items or people and general instances of locations and events, reflecting the Panofsky-Shatford mode/facet matrix of specific, generic, and abstract subjects of pictures.

For TRECVID interactive search experiments to achieve greater ecological validity, the subject pools should be people outside of the system research and development group, i.e., "novices" instead of "experts" using our parlance, as the studies overviewed here confirm that novices and experts will use the system differently. Ideally, representatives of the target community can participate, as was done with the analysts. Work with analysts showed that sports topics carry no meaning for this group, and that the metric of MAP at a depth of 1000 shots is also unrealistic.

The value of a common TRECVID benchmark for evaluation helps greatly, but of course "video search" is much broader than the shot-based retrieval from news corpora discussed here. The work reported here assumes the information need is visual and shot-based, and one alternate approach for satisfying such needs is to repeatedly compound the value of text descriptive metadata for the shots through human intervention. This approach is not investigated here, but merits attention. Just as video access is improved through human intervention, video processing and tagging video with descriptive metadata can be enhanced through human computation. The *ESP Game* has shown that people willingly give their time to recreational games that can have built-in capabilities to collect and refine text descriptors for imagery [15]. Tagging systems have become increasingly popular on the Web for people voluntarily adding free text descriptions to image resources, with a published taxonomy of tagging systems available to help inform their analysis and design [7]. An example of tagging systems' success is demonstrated in the huge volume of annotated images, over 150 million, in the Flickr collection. One could imagine posting news shots for open markup and tagging in either social networking or game-like settings, with such text descriptors forming the basis for an expanded query-by-text functionality, perhaps enhanced enough to obviate the need for other query mechanisms, but such an investigation is beyond the scope of this paper.

The TRECVID 2005 topic sessions provided quantitative and qualitative metrics supporting the interface design as productive for shot-based retrieval tasks by analysts given an expressed information need, the TRECVID topic. Analyst activity is creative and exploratory as well, where the information need is discovered and evolves over time based on interplay with data sources. Likewise, video search activity can be creative and exploratory where the information need is discovered and evolves over time. Evaluating tools for exploratory, creative work is difficult, as acknowledged by Shneiderman and Plaisant [11]. TRECVID may very well broaden its scope to cover other issues in video search, for example the exploratory discovery of materials in video corpora rather than seeking relevant materials for a known, expressed need. The assessment strategies should broaden as well, embracing the use of "Multi-dimensional In-depth Long-term Case-studies (MILC)" [11]. Ideally, MILC research could be conducted with representatives of a real user community over time, to see changing patterns of use and utility as the people gain familiarity and experience with the system. In the term "Multi-dimensional In-depth Long-term Case studies" the multi-dimensional aspect refers to using observations, interviews, surveys, as well as automated logging to assess user performance and interface efficacy and utility. The in-depth aspect is the intense engagement of the researchers with the real users to the point of becoming a partner or assistant. Long-term refers to longitudinal studies that begin with training in use of a specific tool through proficient usage that leads to strategy changes for the expert users. Case studies refer to the detailed reporting about a small number of individuals working on their own problems, in their normal environment. Longitudinal studies have been carried out in HCI and in some information visualization projects, but MILC proposes to refine the methods and expand their scope [11]. An open question is how far video search system researchers can go in measuring the utility of their tools by the success achieved by the users they are studying, i.e., a way to keep technical developments in synergy with human needs.

7. CONCLUSION

Both experts and novices have achieved relative high information retrieval performance on interactive video search tasks compared to fully automated search in TRECVID evaluations through the years. Three user studies are presented here to show an evolution of query activity in novices. In 2004, a system's text search capability was used almost exclusively by university students, even when the text search was severely restricted to only on-screen text and not speech transcription. In 2005 with a redesigned system interface emphasizing query-by-text, query-by-image, and query-by-concept equally, university students made use of all three strategies. A 2006 study, reported in detail here for the first time, confirmed that all three strategies are used by a real-world community, i.e., intelligence analysts mining open source information channels. The 2006 study also confirmed through a within-subjects study that analysts strongly prefer to have all three means of querying rather than a system with just query-by-text, and that the full system with all three query capabilities is easier to learn and easier to use. The analysts perform significantly better with such a complete system versus having only a text query capability. The result is interesting in that these analysts were very proficient in text search systems and strategies (and hence one might have expected a bias toward query-by-text).

Through the brief time period reviewed here, query-by-concept has grown in utility and has great potential to play an even greater role in the future of video search, with the risk that the interface will become too complex as the number of concepts grows from tens to a thousand. Design choices can keep the interface simple with scaffolding to more complex operations where appropriate. Through HCI evaluations that test user reactions to interface design choices and performance with the system as reported here, the latest techniques for video indexing and retrieval can be embraced to better the experiences and outcomes of various video search communities.

8. ACKNOWLEDGMENTS

This work is supported by the National Science Foundation under Grant No. IIS-0205219. Our thanks to NIST and the TRECVID organizers for enabling this video retrieval evaluation work, and to the analysts who participated in this study.

9. REFERENCES

[1] Christel, M.G., and Conescu, R.M. Addressing the Challenge of Visual Information Access from Digital Image and Video Libraries. In *Proc. JCDL* (Denver, CO, June 2005), 69-78.

[2] Christel, M.G., and Conescu, R.M. Mining Novice User Activity with TRECVID Interactive Retrieval Tasks. In *LNCS 4071: Proc. Image and Video Retrieval (CIVR)* (Tempe, AZ, July 2006), Springer, Berlin, 2006, 21-30.

[3] Christel, M., and Moraveji, N. Finding the Right Shots: Assessing Usability and Performance of a Digital Video Library Interface. In *Proc. ACM Multimedia* (New York, NY, Oct. 2004), ACM Press, New York, 2004, 732-739.

[4] Girgensohn, A., Adcock, J., Cooper, M., and Wilcox, L. Interactive Search in Large Video Collections. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, ACM Press, New York, NY, 2005, 1395-1398.

[5] Hauptmann, A.G., and Christel, M.G. Successful Approaches in the TREC Video Retrieval Evaluations. In *Proc. ACM Multimedia* (Oct. 2004), 668-675.

[6] Hollink, L., Nguyen, G.P., Koelma, D.C., Schreiber, A.T., and Worring, M. Assessing User Behaviour in News Video Retrieval. *IEE Proc. Vision, Image, & Signal Processing 152*(6), 2005, 911-918.

[7] Marlow, C., Naaman, M., Boyd, D., and Davis, M. HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, To Read. In *Proc. ACM Hypertext and Hypermedia* (Odense, Denmark, Aug. 2006), ACM Press, New York, NY, 31-40.

[8] Naphade, M., Smith, J. R., Tesic, J., Chang, S.-F., Hsu, W., Kennedy, L., Hauptmann, A. and Curtis, J. Large-Scale Concept Ontology for Multimedia. *IEEE MultiMedia 13*(3), 2006, 86-91.

[9] National Institute of Standards and Technology, NIST TREC Video Retrieval Evaluation Online Proceedings, 2001-2006, http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html.

[10] Shneiderman, B., Byrd, D., and Croft, W.B. Clarifying Search: A User-Interface Framework for Text Searches. *D-Lib Magazine, 3*, 1 (January 1997), http://www.dlib.org.

[11] Shneiderman, B., and Plaisant, C. Strategies for Evaluating Information Visualization Tools: Multi-dimensional In-depth Long-term Case Studies. In *Proc. BELIV'06 Workshop, Advanced Visual Interfaces Conf.* (Venice, May 2006), 1-7.

[12] Smeulders, A., Worring, M., Santini, S., Gupta, A., and Jain, R. Content Based Image Retrieval at the End of the Early Years. *IEEE Trans. Pattern Analysis and Machine Intelligence 22*(12), 2000, 1349-1380.

[13] Snoek, C., Worring, M., Koelma, D., and Smeulders, A. Learned Lexicon-Driven Interactive Video Retrieval. In *LNCS 4071: Proc. Image and Video Retrieval (CIVR)* (Tempe, AZ, July 2006), Springer, Berlin, 2006, 11-20.

[14] Snoek, C., Worring, M., Koelma, D., and Smeulders, A. A Learned Lexicon-Driven Paradigm for Interactive Video Retrieval. *IEEE Trans. Multimedia 9*(2), Feb. 2007, 280-292.

[15] von Ahn, L., and Dabbish, L. Labeling Images with a Computer Game. In *Proc. ACM CHI* (Vienna, Austria, April 2004), ACM Press, New York, NY, 2004, 319-326.

[16] Yan, R., and Hauptmann, A.G. Efficient Margin-Based Rank Learning Algorithms for Information Retrieval. In *LNCS 4071: Proc. Image and Video Retrieval (CIVR)* (Tempe, AZ, July 2006), Springer, Berlin, 2006, 113-122.

[17] Yang, M., Wildemuth, B., and Marchionini, G. The Relative Effectiveness of Concept-based Versus Content-based Video Retrieval. In *Proc. ACM Multimedia* (Oct. 2004), 368-371.